Improving State-of-the-Art Continuous Speech Recognition Systems Using the N-Best Paradigm with Neural Networks

S. Austin, G. Zavaliagkos†, J. Makhoul, and R. Schwartz

BBN Systems and Technologies, Cambridge, MA 02138
†Northeastern University, Boston, MA 02115

ABSTRACT

In an effort to advance the state of the art in continuous speech recognition employing hidden Markov models (HMM), Segmental Neural Nets (SNN) were introduced recently to ameliorate the well-known limitations of HMMs, namely, the conditional-independence limitation and the relative difficulty with which HMMs can handle segmental features. We describe a hybrid SNN/HMM system that combines the speed and performance of our HMM system with the segmental modeling capabilities of SNNs. The integration of the two acoustic modeling techniques is achieved successfully via the N-best rescoring paradigm. The N-best lists are used not only for recognition, but also during training. This discriminative training using N-best is demonstrated to improve performance. When tested on the DARPA Resource Management speaker-independent corpus, the hybrid SNN/HMM system decreases the error by about 20% compared to the state-of-the-art HMM system.

INTRODUCTION

In February 1991, we introduced at the DARPA Speech and Natural Language Workshop the concept of a Segmental Neural Net (SNN) for phonetic modeling in continuous speech recognition [1]. The SNN was introduced to overcome some of the well-known limitations of hidden Markov models (HMM) which now represent the state of the art in continuous speech recognition (CSR). Two such limitations are (i) the conditional-independence assumption, which prevents a HMM from taking full advantage of the correlation that exists among the frames of a phonetic segment, and (ii) the awkwardness with which segmental features (such as duration) can be incorporated into HMM systems. We developed the concept of SNN specifically to overcome the two HMM limitations just mentioned for phonetic modeling in speech. However, neural nets are known to require a large amount of computation, especially for training. Also, there is no known efficient search technique for finding the best scoring segmentation with neural nets in continuous speech. Therefore, we have developed a hybrid SNN/HMM system that is designed to take full advantage of the good properties of both methods: the phonetic modeling properties of SNNs and the good computational properties of HMMs. The two methods are integrated through the use of the N-best paradigm, which was developed in conjunction with the BYBLOS system at BBN [7,6].

A year ago, we presented very preliminary results using our hybrid system on the speaker-dependent portion of the DARPA Resource Management Corpus [1]. Also, the training of the neural net was performed only on the correct transcription of the utterances. In this paper, we describe the performance of the hybrid system on the speaker-independent portion of the Resource Management corpus, using discriminative training on the whole N-best list. Below, we give a description of the SNN, the integration of the SNN with the HMM models using the N-best paradigm, the training of the hybrid SNN/HMM system using the whole N-best list, and the results on a development set.

SEGMENTAL NEURAL NET STRUCTURE

The SNN differs from other approaches to the use of neural networks in speech recognition in that it attempts to recognize each phoneme by using all the frames in a phonetic segment simultaneously to perform the recognition. The SNN is a neural network that takes the frames of a phonetic segment as input and produces as output an estimate of the probability of a phoneme given the input segment. But the SNN requires the availability of some form of phonetic segmentation of the speech. To consider all possible segmentations of the input speech would be computationally prohibitive. We describe in Section 3 how we use the HMM to obtain likely candidate segmentations. Here, we shall assume that a phonetic segmentation has been made available.

The structure of a typical SNN is shown in Figure1. The input to the net is a fixed number of frames of speech features (5 frames in our system). The features in each 10-ms frame consist of 16 scalar values: 14 mel-warped cepstral coefficients, power, and power difference. Thus, the input to the SNN consists of a total of 80 features. But the actual number of actual frames in a phonetic segment is variable. Therefore, we convert the variable number of frames in each segment to a fixed number of frames (in this case, five frames). In this way, the SNN is able to deal effectively with variable-length segments in continuous speech. The requisite time warping is performed by a quasi-linear sampling of the feature vectors comprising the segment. For example, in a 17-frame phonetic segment, we would use frames 1, 5, 9, 13, and 17 as input to the SNN. In a 3-frame segment, the five frames used are 1, 1, 2, 3, 3, with a repetition of the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1992		**00-00-1992 to 00-00-1992**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Improving State-of-the-Art Continuous Speech Recognition System Using the N-Best Paradigm with Neural Networks	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **BBN Technologies,10 Moulton Street,Cambridge,MA,02238**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**5**	

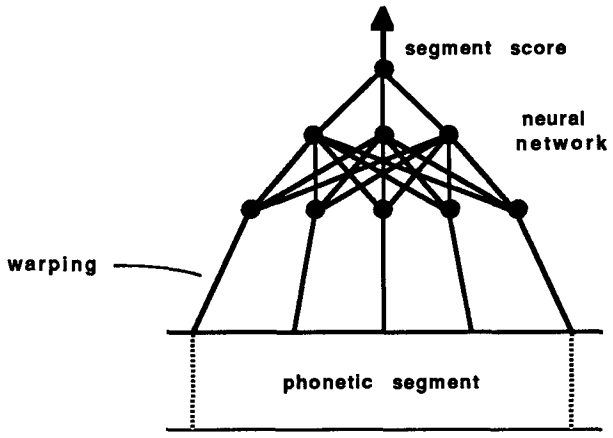

Figure 1: The Segmental Neural Network model samples the frames in a segment and produces a single segment score.

first and third frames. In this sampling, we are using a result from Stochastic Segment Models (SSM) [5] in which it was found that sampling of naturally-occurring frames gives better results than strict linear interpolation.

Since there are 53 phonemes defined in our system, we used SNNs with 53 outputs, each representing one of the phonemes in the system.

THE N-BEST RESCORING PARADIGM

Without an algorithm that can efficiently search all word-sequence and segmentation posibilities in a large-vocabulary CSR system, the amount of computation required to incorporate the SNN into such a system would be prohibitive. However, it is possible to use the N-best paradigm to make such an incorporation feasible.

The N-best paradigm [7,6] was originally developed at BBN as a simple way to ameliorate the effects of errors in speech recognition when integrating speech with natural language processing. Instead of producing a single sequence words, the recognition component produces a list of N best-scoring sequences. %The list of N sentences is ordered by overall score in matching the input utterance. For integration with natural language, we send the list of N sentences to the natural language component, which processes the sentences in the order given and chooses the highest scoring sentence that can be understood by the system. However, we found that the N-best paradigm can also be very useful for improving speech recognition performance when more expensive sources of knowledge (such as cross-word effects and higher-order statistical grammars) cannot be computed efficiently during the recognition. All one does is rescore the N-best list with the new sources of knowledge and reorder the list. The SNN is a good example of an expensive knowledge source, whose use would benefit greatly from using N-best rescoring, thus comprising a hybrid SNN/HMM system.

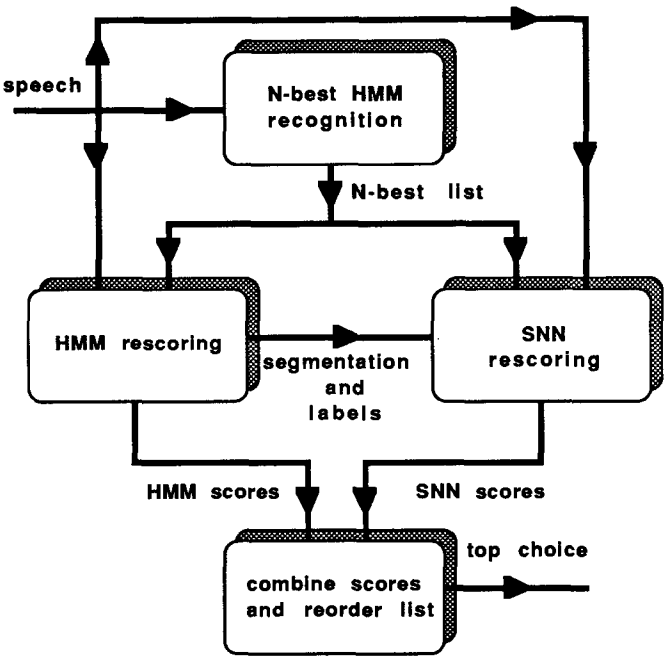

Figure 2: Schematic diagram of the hybrid SNN/HMM system using the N-best rescoring paradigm.

Figure 2 shows a block diagram of the hybrid SNN/HMM system. A spoken utterance is processed by the HMM recognizer to produce a list of the N best-scoring sentence hypotheses. The length of this list is chosen to be long enough to almost always include the correct answer (from experience, N=20 is usually sufficient). Thereafter, the recognition task is reduced to selecting the best hypothesis from the N-best list. Because these N-best lists are quite short (e.g., N=20), each hypothesis can be examined and scored using algorithms which would have been computationally impossible with a search through the entire vocabulary. In addition, it is possible to generate several types of scoring for each hypothesis. This not only provides a very effective means of comparing the effectiveness of different speech models (e.g., SNN versus HMM), but it also provides an easy way to combine several radically different models.

One most obvious way in which the SNN could use the N-best list would be to use the HMM system to generate a segmentation for each N-best hypothesis (by finding the most likely HMM state sequence according to that hypothesis) and to use the SNN to generate a score for the hypothesis using this segmentation. This SNN score for a hypothesis is the logarithm of the product of the appropriate SNN outputs for all the segments in a segmentation according to that hypothesis. The chosen answer would be the hypothesis with the best SNN score. However, it is also possible to

generate several scores for each hypothesis, such as SNN score, HMM score (which is the logarithm of the HMM likelihood), grammar score, and the hypothesized number of words and phonemes. We can then generate a composite score by, for example, taking a linear combination of the individual scores. After we have rescored the N-Best list, we can reorder it according to the new scores. If the CSR system is required to output just a single hypothesis, the highest scoring hypothesis is chosen. We call this whole process *the N-best rescoring paradigm.*

The linear combination that comprises the composite score is determined by selecting the weights that give the best performance over a development test set. These weights can be chosen automatically [4]. The number of words and phonemes are included in the composite score because they serve the same purpose as word and phoneme insertion penalties in a HMM CSR system.

SEGMENTAL NEURAL NET TRAINING

1-Best Training

In our original training algorithm, we first segmented all of the training utterances into phonetic segments using the HMM models and the utterance transcriptions. Each segment then serves as a positive example of the SNN output corresponding to the phonetic label of the segment and as a negative example for all the other 52 phonetic SNN outputs. We call this training method *1-best training.*

The SNN was originally trained using a mean-square error (MSE) criterion – i.e., the SNN was trained to minimize

$$E = \frac{1}{N}\sum_{n=1}^{N}(y_c(n) - d_c(n))^2$$

where $y_c(n)$ is the network output for phoneme class c for the n^{th} training vector and $d_c(n)$ is the desired output for that vector (1 if the segment belongs to class c and 0 otherwise). This measure can lead to gross errors at low values of $y_c(n)$ when segment scores are multiplied together. Accordingly, we adopted the log-error training criterion [3], which is of the form

$$E = -\frac{1}{N}\sum_{n=1}^{N}\log\left(y_c(n) - [1 - d_c(n)]\right)^2.$$

This can be shown to have several advantages over the MSE criterion. When the neural net non-linearity is the usual sigmoid function, this error measure has only one minimum for single layer nets. In addition, the gradient is simple and avoids the problem of "weight locking" (where large errors do not change because of small gradients in the sigmoid).

Duration

Because of the time-warping function (which transforms phonetic segments of any length into a fixed-length representation), the SNN score for a segment is independent of the duration of the segment. In order to provide information about the duration to the SNN, we constructed a simple durational model. For each phoneme, a histogram was made of segment durations in the training data. This histogram was then smoothed by convolving with a triangular window, and probabilities falling below a floor level were reset to that level. The duration score was multiplied by the neural net score to give an overall segment score.

N-best Training

In our latest version of the training algorithm, we take the N-best paradigm a step further and perform what we call *N-best training*, which is a form of discriminative training. First, we take the HMM-based segmentations of the training utterances according to the correct word sequence. These segments are used only as positive examples (i.e., trained to output 1) for the appropriate SNN outputs.

We then produce the N-best lists for all of the training sentences. For each of the incorrect hypotheses in the N-best list, we obtain the HMM-based segmentation and isolate those segments that differ from the segmentation according to the correct transcription and use them as negative training for the SNN outputs (i.e., trained to output 0). Thus we train negatively on the "misrecognized" parts of the incorrect hypothesis.

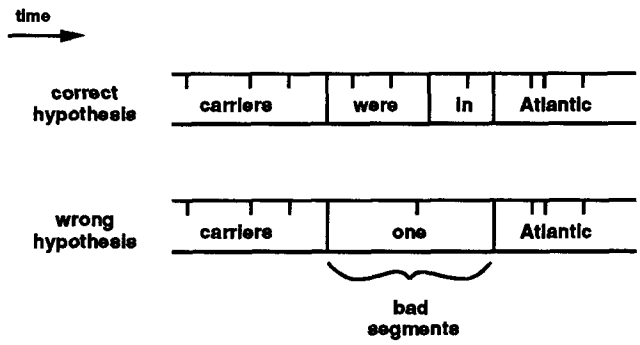

Figure 3: N-Best training trains the SNN to specifically reject those segments from an incorrect hypothesis that the HMM consioders likely

This new method has the advantage that the SNN is specifically trained to discriminate among the choices that the HMM system considers difficult. This is better than the 1-best training algorithm, which only uses the segmentation of the correct utterance transcription, because N-best training directly optimizes the performance of the SNN in the N-best rescoring paradigm.

If, for example, the transcription of part of an utterance "...carriers were in Atlantic..." and a likely N-best hypothesis was "...carriers one Atlantic..." the segments corresponding to the word "one" (as generated by a constrained HMM alignment) would be presented to the SNN as nega-

Table 1: SNN development on February '89 test set

		Word Error (%)
	Original SSN (MSE)	13.7
+	Duration	12.7
+	Log-Error Criterion	11.6
+	N-Best training	9.0

Table 2: Hybrid SNN/HMM system: test results.

System*	*N*	Feb '89	Oct '89
HMM	1	3.5	3.8
SNN	20	9.0	—
SNN+HMM	2	3.3	—
SNN+HMM	4	2.9	—
SNN+HMM	20	2.8	3.0

* All systems include word and segment scores.

tive training. To determine if a segment should be presented to the SNN as negative input, the label and position of each segment in a hypothesis is compared to the segments seen in the correct segmentation of the utterance. If either the label or the position (subject to a tolerance threshold) of the segment does not match a segment in the correct segmentation, it is presented as negative training.

EXPERIMENTAL CONDITIONS AND RESULTS

Experiments to test the performance of the hybrid SNN/HMM system were performed on the Speaker Independent (SI) portion of the DARPA 1000-word Resource Management speech corpus, using the standard word-pair grammar (perplexity 60). The training set consisted of utterances from 109 speakers, 2830 utterances from male speakers and 1160 utterances from female speakers, and the February '89 test set was used for development of the system. The October '89 test set was used for the final independent test.

In our initial experiments, we used the February '89 development set. Table 1 shows the word error rates when we rescored the N=20 N-best lists at the various stages of development of the SNN. It should be noted that the figures do not reflect the unaided performance of the SNN in recognition, since the N-best list was generated by a HMM system, but instead illustrate the effectiveness of the respective improvements.

The original 1-layer SNN was trained using the 1-best training algorithm and the MSE criterion; it gave an error rate of 13.7%. The incorporation of the duration term and the adoption of the log-error training criterion both resulted in some improvement, bringing the error rate to 11.6%.

When we used the N-best training (which used the SNN produced by the 1-best training as an initial estimate), the error rate dropped to 9.0%, confirming our belief that the N-best training is more effective than the 1-best training in the N-best rescoring paradigm. This final condition was then used to generate the SNN score to examine the behavior of the hybrid SNN/HMM system.

Table 2 shows the results of combining the HMM and SNN scores in the re-ordering of the N-Best list. Taking the top answer of the N-best list (as produced by the HMM system) gave an error rate of 3.5% on the February '89 development test set. Upon re-ordering the N=20 list on the basis of the SNN score alone, the error rate was 9.0%. However, upon combining the HMM and SNN scores, the error rate decreased over that of the HMM alone. The error rate decreased as the value of N used in the N-best list was increased. For N=2, the error decreased to 3.3%, then to 2.9% for N=4, and finally to 2.8% for N=20.

Based upon the results of the February'89 development set, we rescored the 20-best lists generated from the October '89 with the hybrid system. This independent test yielded an even larger improvement, reducing the error rate from 3.8% in the HMM-based system to 3.0% in the SNN/HMM system. This represents a 20% reduction in error rate.

Given that the HMM system used in our experiments represented the state of the art in CSR, the hybrid SNN/HMM system has now established a new state of the art.

CONCLUSIONS

We have presented the Segmental Neural Net as a method for phonetic modeling in large vocabulary CSR systems and have demonstrated that, when combined with a conventional HMM, the SNN gives an improvement over the performance of a state-of-the-art HMM CSR system.

We have used the N-best rescoring paradigm to achieve this improvement in two ways. Firstly, the N-best rescoring paradigm has allowed us to design and test the SNN with little regard to the usual problem of searching when dealing with a large vocabulary speech recognition system. Secondly, the paradigm provides a simple way of combining the best aspects of two systems, leading to a combined system which exceeds the performance of either one alone.

Future work will concentrate on modifying the N-best training algorithm to model context in the SNN. We will also investigate possible improvements to the structure of the SNN, including different network architectures and additional segment features.

Acknowledgments

The authors would like to thank Amro El-Jaroudi of the University of Pittsburgh for his help in several aspects of this work. This work was sponsored by DARPA.

REFERENCES

1. Austin, S., Makhoul, J., Schwartz, R., and Zavaliagkos, G., "Continuous Speech Recognition Using Segmental Neural Nets," *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, pp. 249-252, Morgan Kaufmann Publishers, February 1991.
2. Austin, S., Peterson, P., Placeway, P., Schwartz, R., Vandegrift, J., "Towards a Real-Time Spoken Language System Using Commercial Hardware," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, Morgan Kaufmann Publishers, June 1990.
3. El-Jaroudi, A. and Makhoul, J., "A New Error Criterion for Posterior Probability Estimation with Neural Nets," *International Joint Conference on Neural Networks*, San Diego, CA, June 1990, Vol III, pp. 185-192.
4. Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R., Rohlicek, J.R., "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, Morgan Kaufmann Publishers, February 1991.
5. Ostendorf, M. and Roukos S., "A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition," *IEEE Trans. Acoustic Speech and Signal Processing*, Vol. ASSP-37(12), December 1989, pp. 1857–1869.
6. Schwartz, R. and Austin, S., "A comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," ICAASP-91, Toronto, Canada, May 1991, pp. 701-704.
7. Schwartz, R. and Chow, Y.L., "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," ICASSP-90, Albuquerque, NM, April 1990, pp. 81-84.